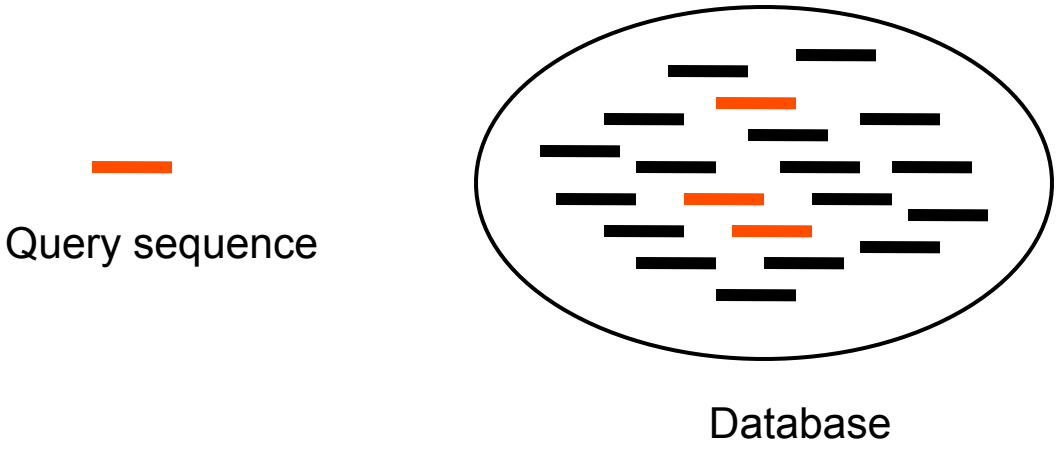
BLAST

Anders Gorm Pedersen

Database searching

Using pairwise alignments to search databases for similar sequences



Database searching

Most common use of pairwise sequence alignments is to search databases for related sequences. For instance: find probable function of newly isolated protein by identifying similar proteins with known function.

Most often, ***local*** alignment (“Smith-Waterman”) is used for database searching: you are interested in finding out if ANY domain in your protein looks like something that is known.

Often, full Smith-Waterman is too time-consuming for searching large databases, so heuristic methods are used (fasta, BLAST).

BLAST: about 100 x faster than full Smith-Waterman

BLAST flavors

BLASTN
- Nucleotide query sequence
- Nucleotide database

BLASTP
- Protein query sequence
- Protein database

BLASTX
- Nucleotide query sequence
- Protein database
- Compares all six reading frames with the database

TBLASTN
- Protein query sequence
- Nucleotide database
- ”On the fly” six frame translation of database

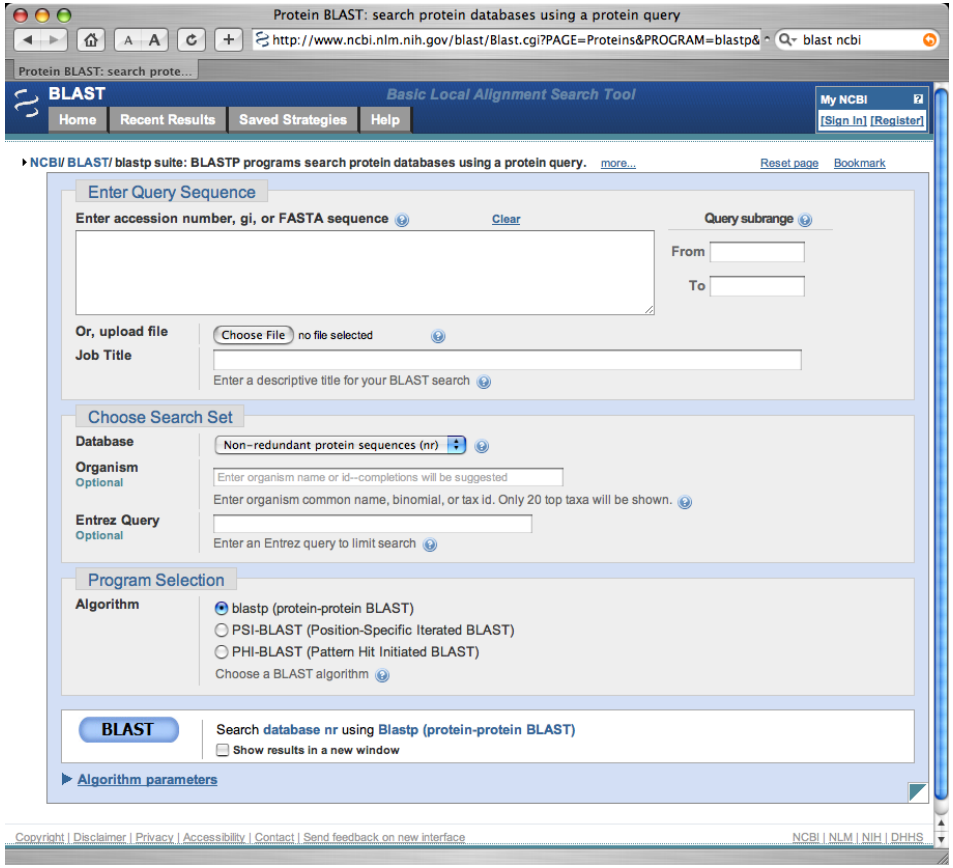
TBLASTX
- Nucleotide query sequence
- Nucleotide database
- Compares all reading frames of query with all reading frames of the database

Searching on the web: BLAST at NCBI

Very fast computer dedicated to running BLAST searches

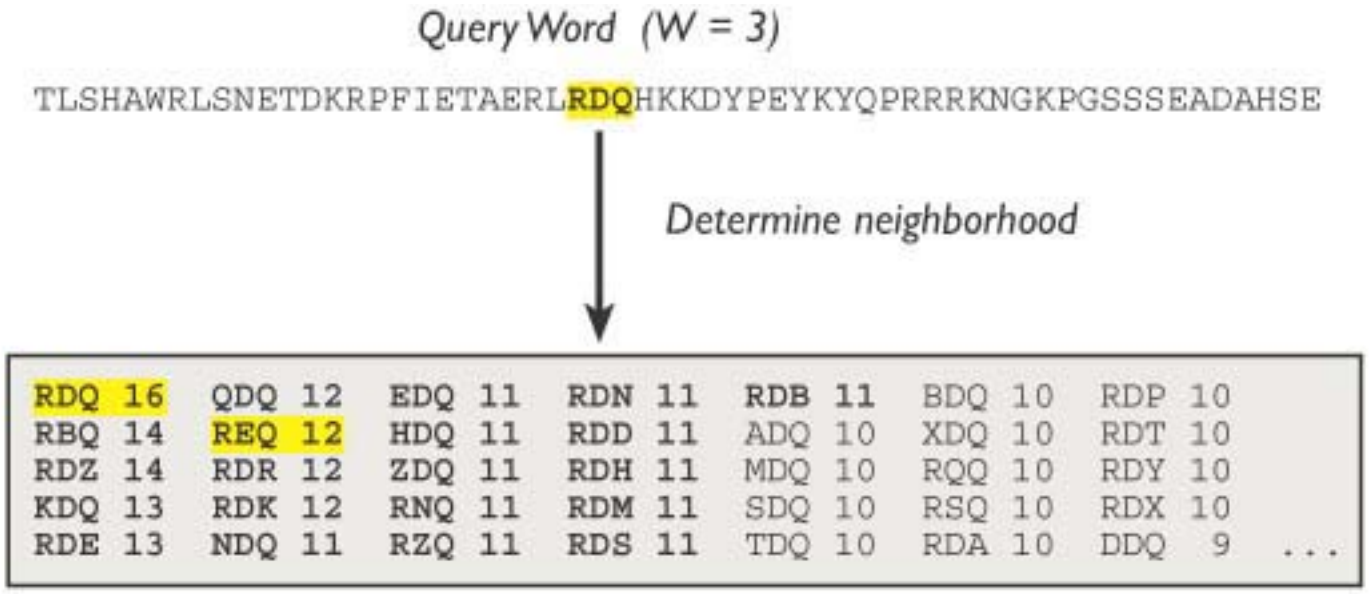
Many databases that are always up to date (e.g. NR and Human Genome)

Nice simple web interface



Why is BLAST faster? (Protein search example)

- Breaks up query sequence into smaller “words” of length 3 aa
- For each word: Determine “neighborhood” = list of other words (of same length) that have a high alignment score to starting word



Why is BLAST faster? (Protein search example)

- For each word in neighborhood: Find exact matches in database sequences
- If two words on same “alignment diagonal” and within 40aa: try to extend alignment in either direction to create optimal local alignment
- Various heuristics for deciding when to stop extension (stop when score drops more than X, trim back to previous maximum)

Figure 5-4. Isolated and clustered words

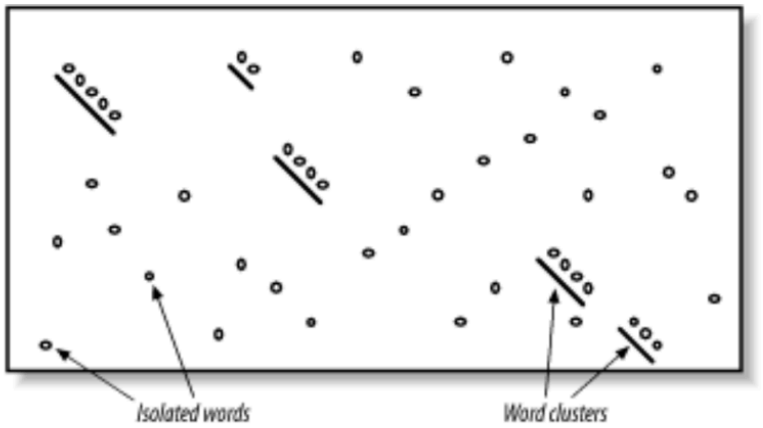
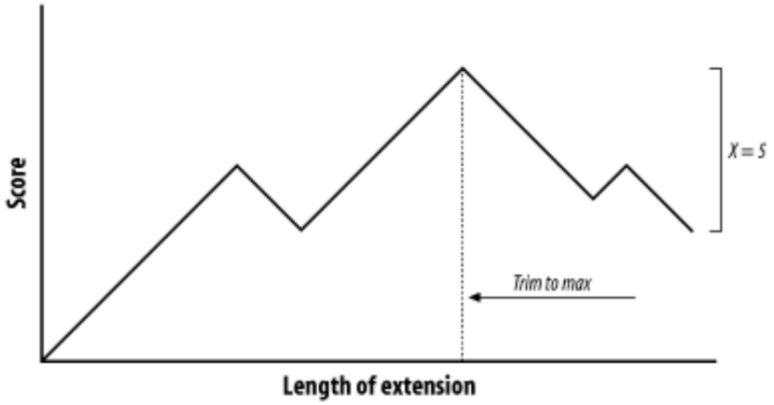


Figure 5-6. Attenuating extension with X



When is a database hit meaningful?

- **Problem**:
  - Even unrelated sequences can be aligned (yielding a low score)
  - How do we know if a database hit is meaningful?
  - When is an alignment score sufficiently high?

- **Solution**:
  - Determine the range of alignment scores you would expect to get for random reasons (i.e., when aligning unrelated sequences).
  - Compare actual scores to the distribution of random scores.
  - Is the real score much higher than you'd expect by chance?

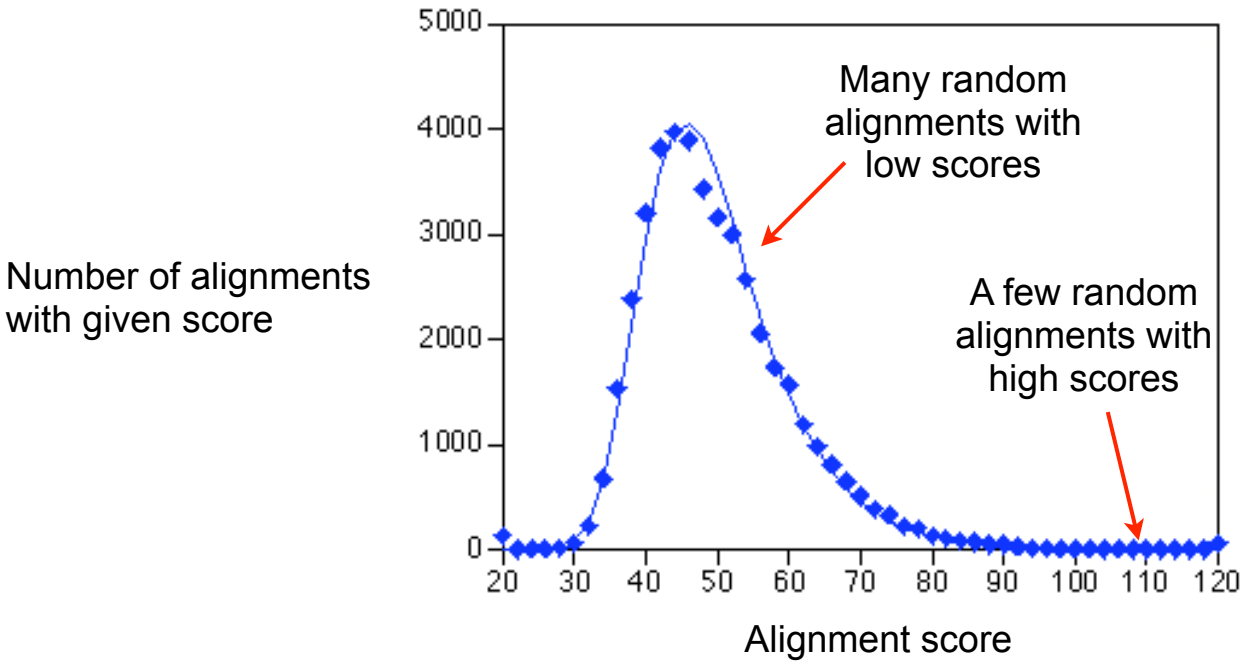
Distribution of random alignment scores

- **Software simulation**

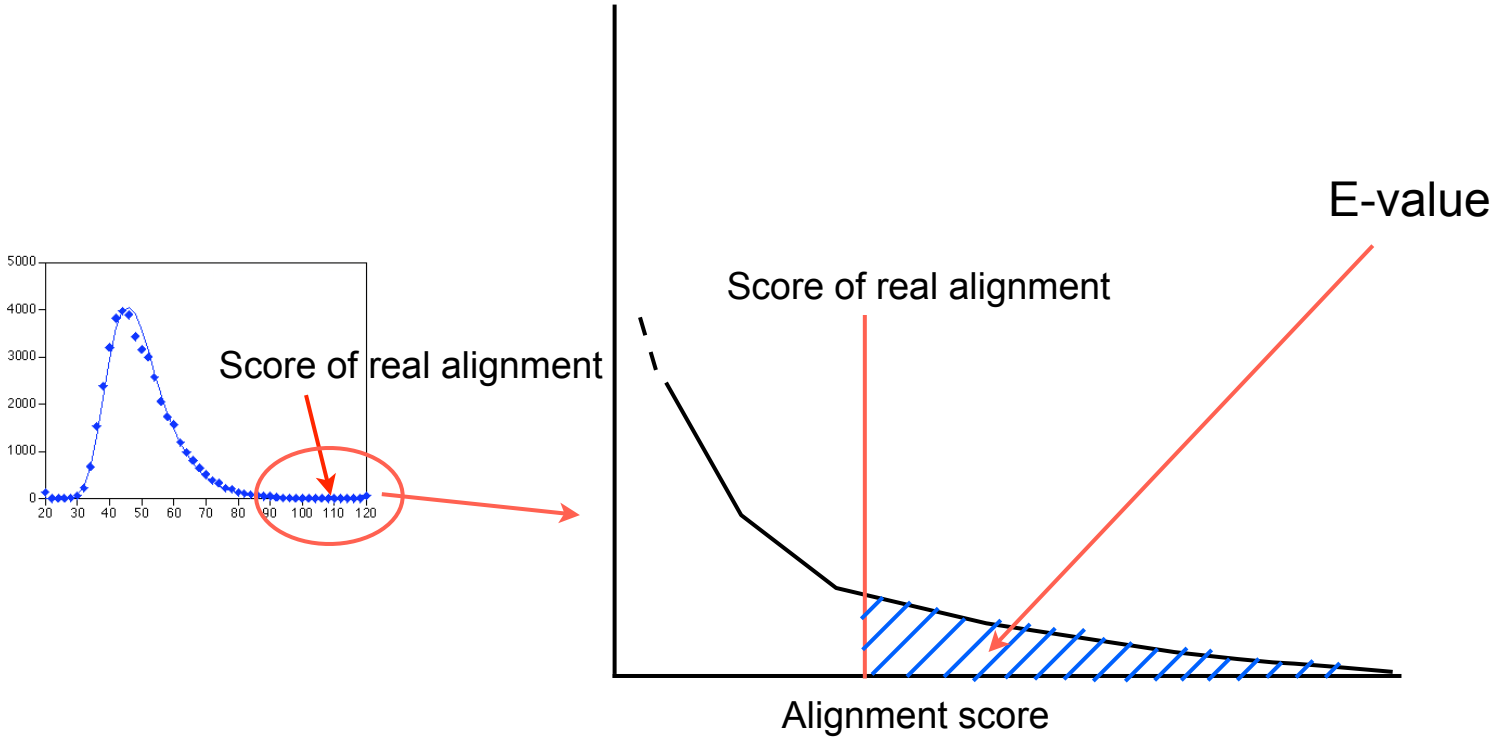
Significance of alignment score expressed as E-value

Pairwise alignment of unrelated sequences results in scores following an extreme value distribution

The exact shape and location of the distribution depends on the length and composition of the sequences



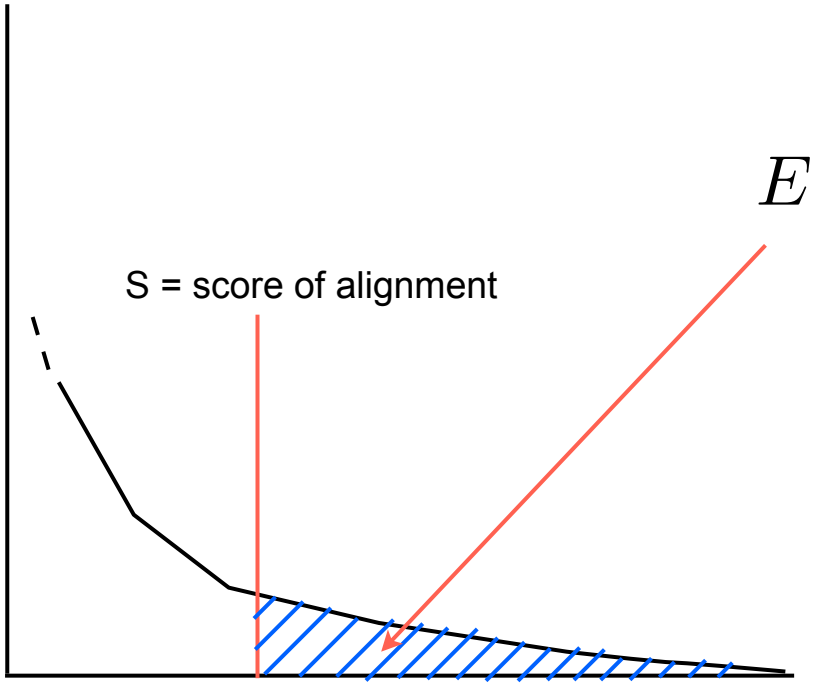
Significance of alignment score expressed as E-value



E-value: the number of random hits with score ≥ real score

Want E-values well below 1 (the lower the better)

Significance of alignment score expressed as E-value



$$E = K \cdot m \cdot n \cdot e^{-\lambda S}$$

m: length of query sequence

n: combined length of all database sequences

λ: the scaling factor we also saw when computing BLOSUM62

K: a constant whose value depends on the nature of the sequences - it can be determined empirically by curve fitting